

Histograms, kernel density estimation, and probability

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 12, Fall 2018

Monday March 4

Plan

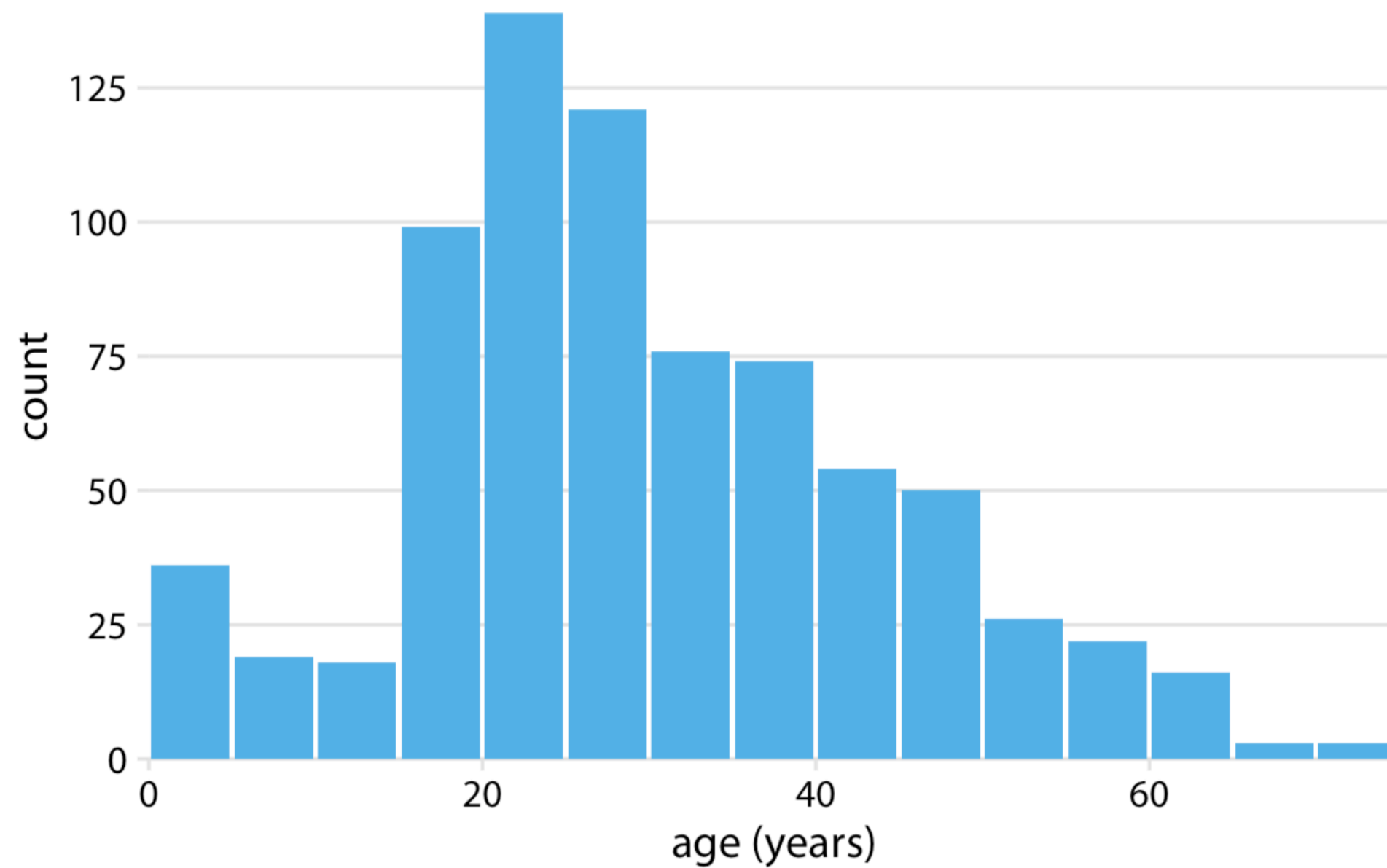
- Histograms
- Kernel density estimation
- Probability and random variables
- Prior probability

Histograms

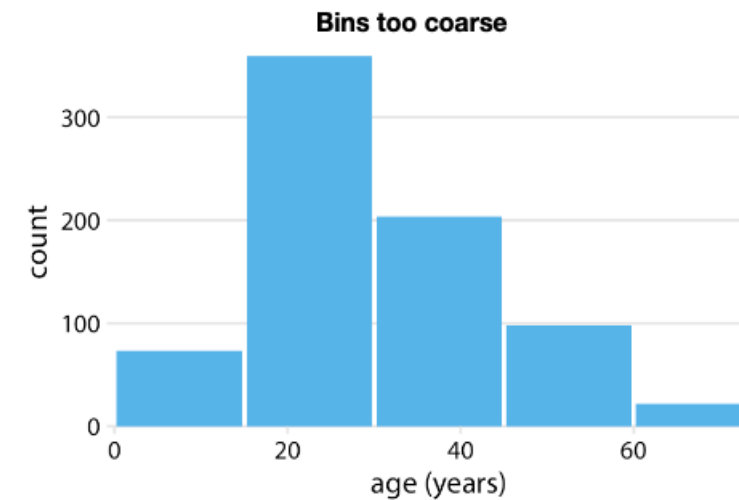
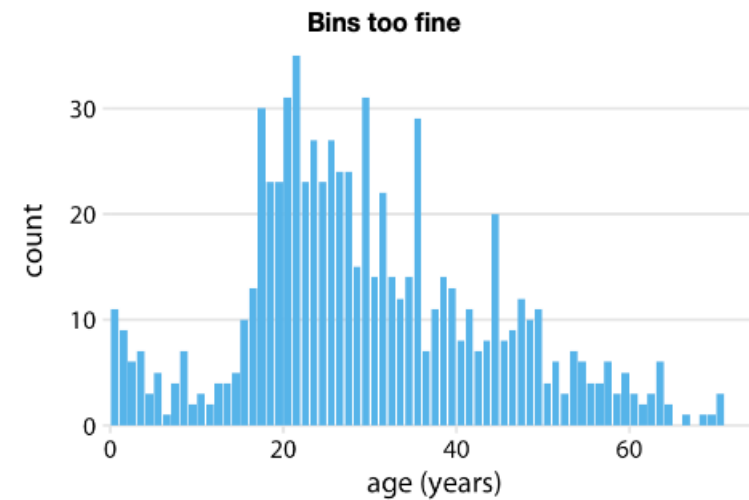
- Visualize counts (or frequency data) as a function of a *discrete*, ordered factor OR quantitative data (numeric ranges)
- For example number of survivors by age on Titanic (Adapted from Wilke, 2019):

Age range	Count	Age range	Count	Age range	Count
0-5	36	31-35	76	61-65	16
6-10	19	36-40	74	66-70	3
11-15	18	41-45	54	71-75	3
16-20	99	46-50	50		
21-25	139	51-55	26		
26-30	121	55-60	22		

Titanic survivor histogram



The binning problem



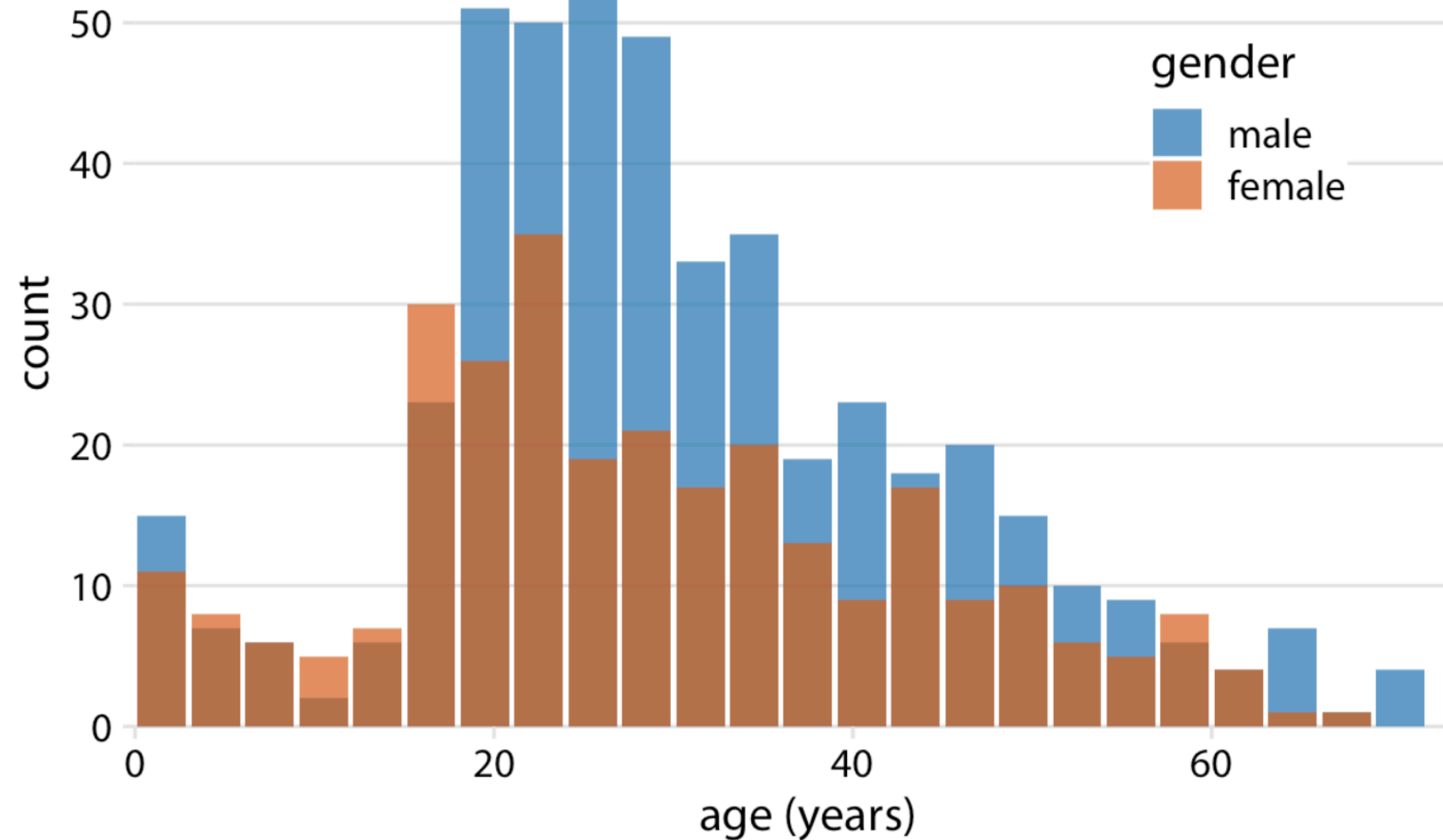
- For discrete, quasi-continuous, and continuous quantitative data we must decide on the interval to lump data into a "bar" (**bin**). i.e. *how many bars/bins do we use to visualize the count data?*
- Too fine: lose big picture to noise
- Too coarse: lose nuance in data

Binning considerations

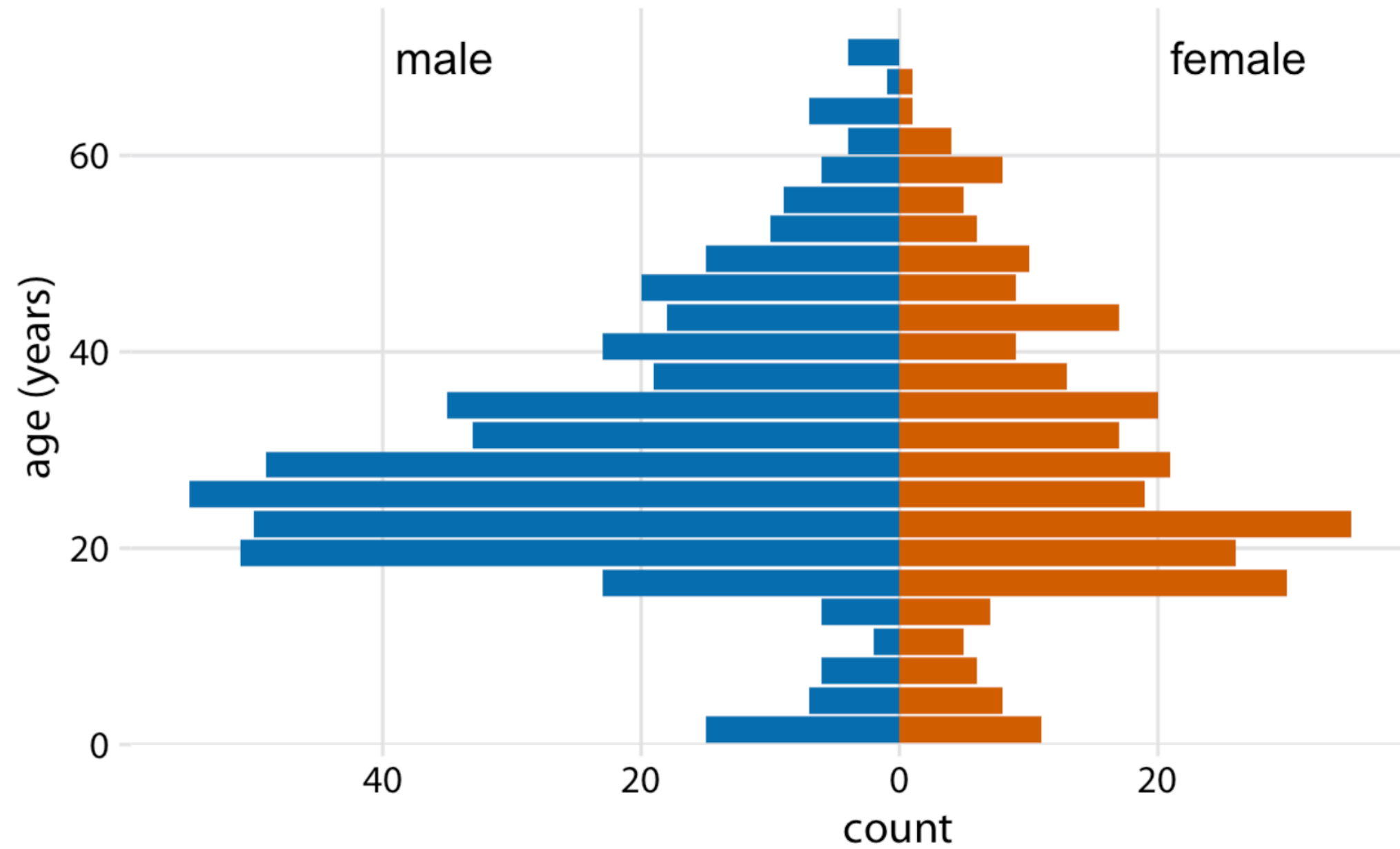
- What is the frequency of real change in the distribution? Bin size should show real changes.
- What is the precision of the measurement?
 - Bin size should not be finer than precision.
- What is the purpose of the histogram? Bin size and range should support the goals.
- Always explore multiple bin widths when making histograms.

Multiple histograms

Do not use stacked bar graphs (why?)



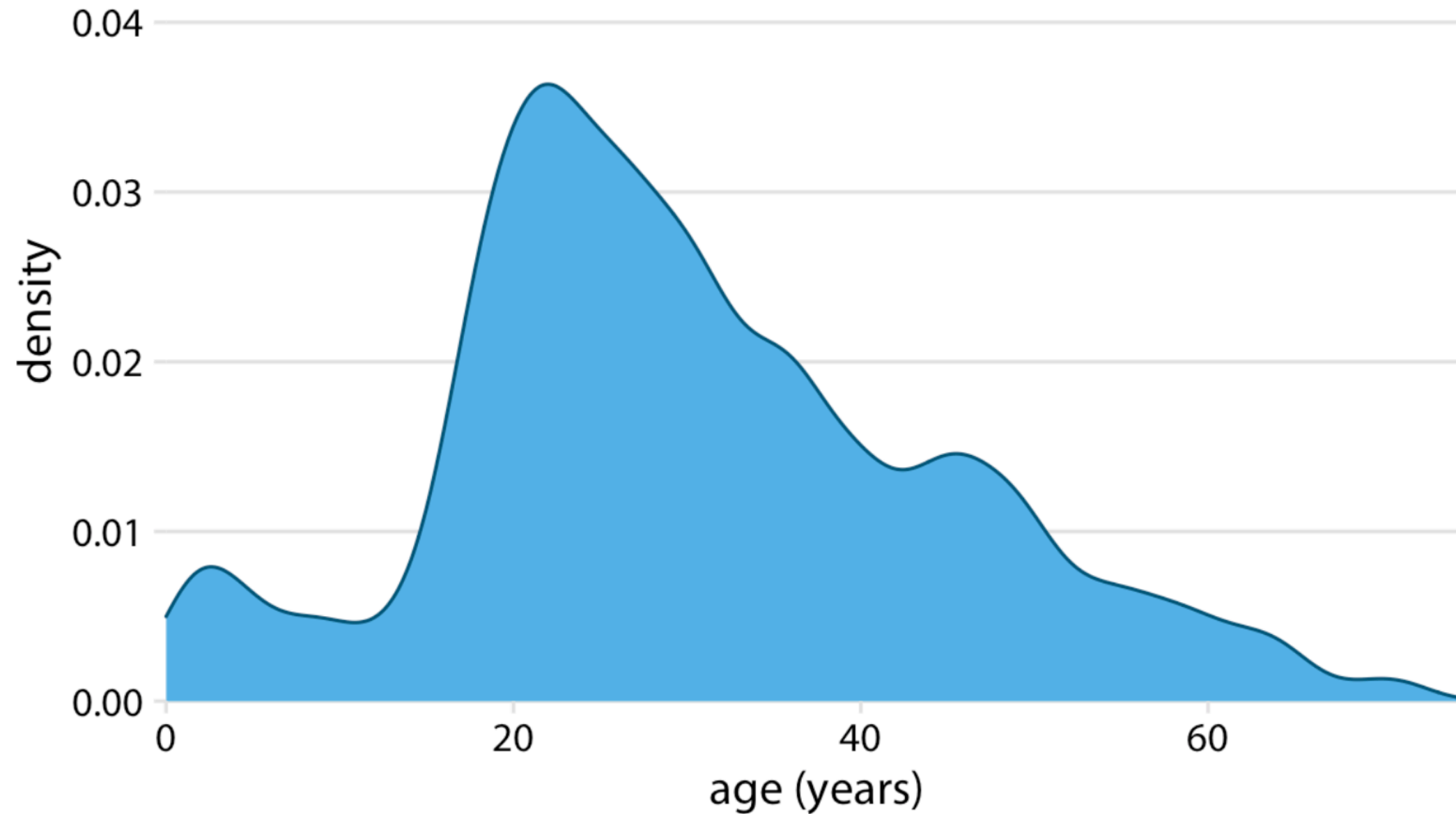
Better approach: "pyramid"



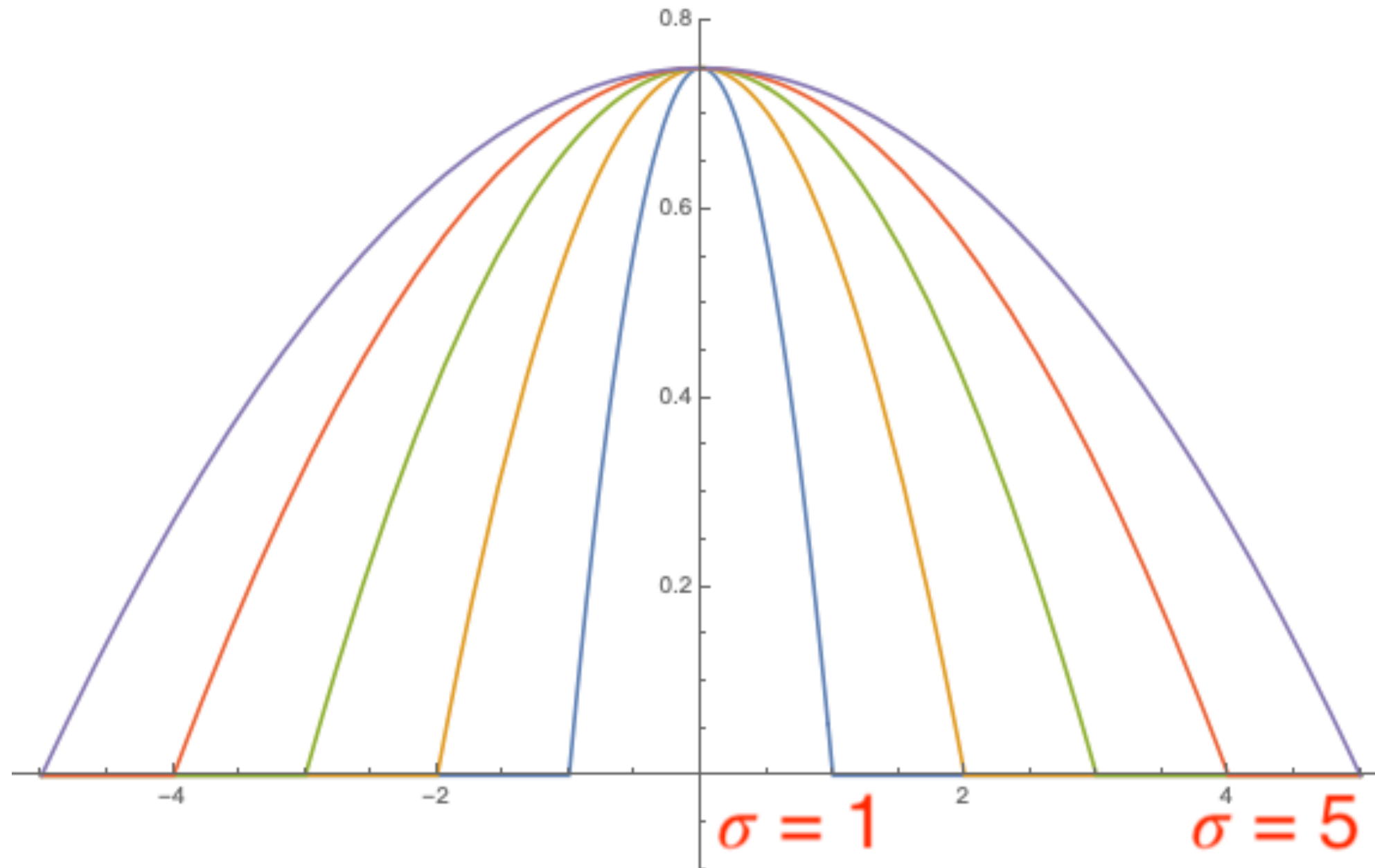
Estimating counts at every x-value

- Scenario: We want estimated survivor count for people age 35 (not interested in range 31-35).
- Don't want finer bins. Why?
 - Bin 35 very sensitive to neighbors in sample.
- We can estimate a continuous curve to get counts for EVERY age we'd like, a process called **kernel density estimation**.
- We normalize y-axis within the data range, so it becomes FREQUENCY rather than COUNT.
 - Often y-axis called **density**: counts ("mass") per unit space (x).

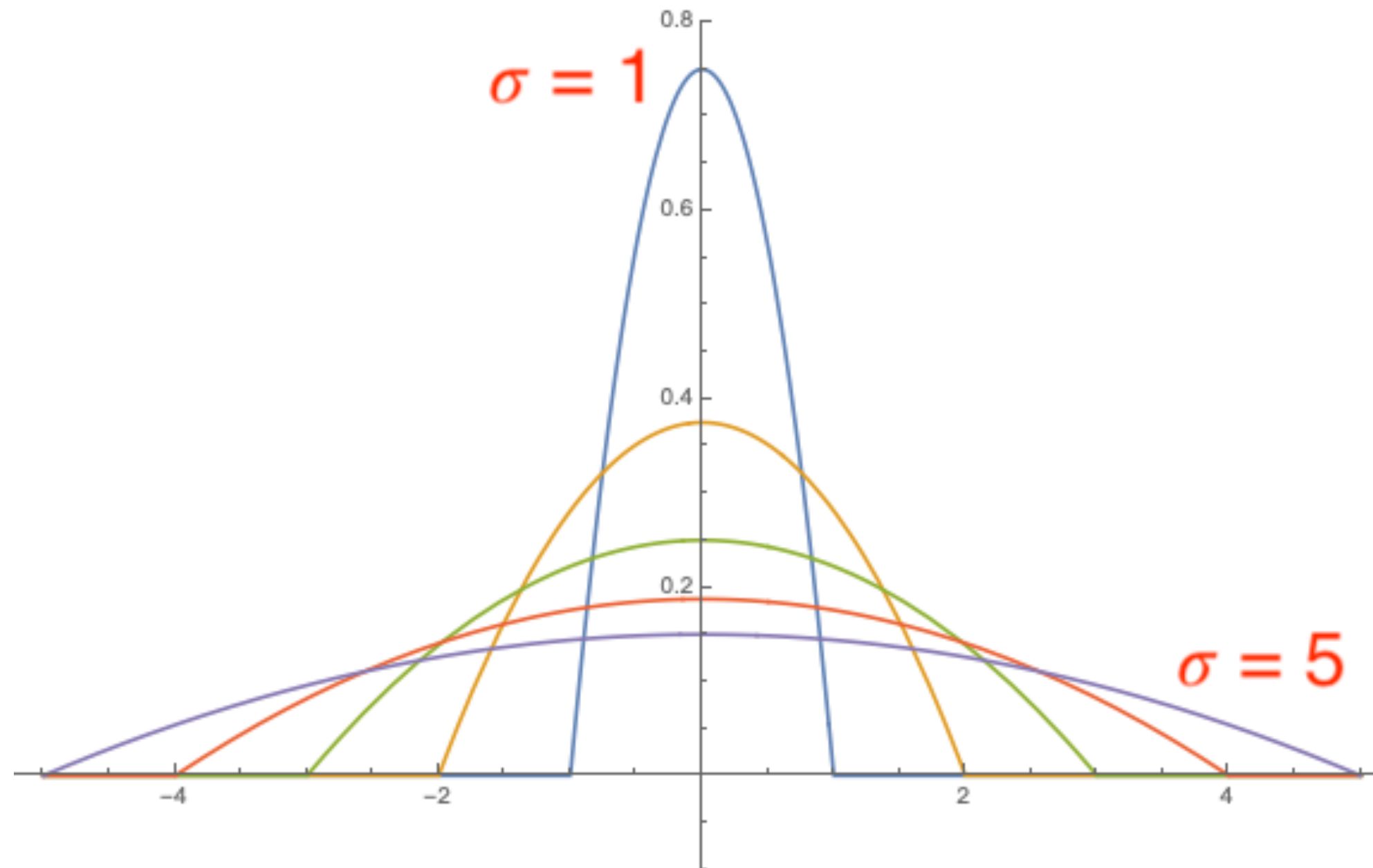
Kernel density estimate (KDE)



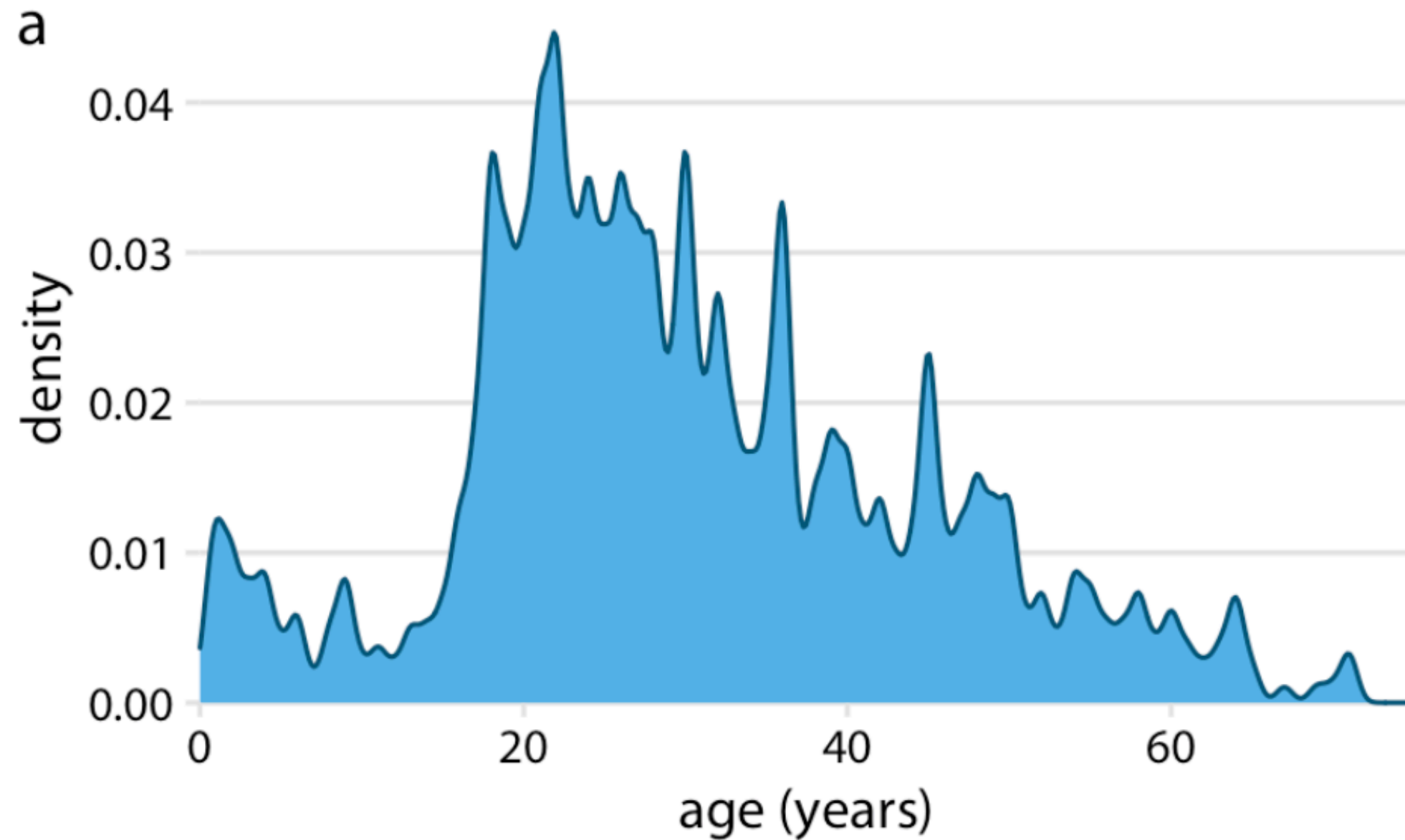
Different parabolic kernel widths



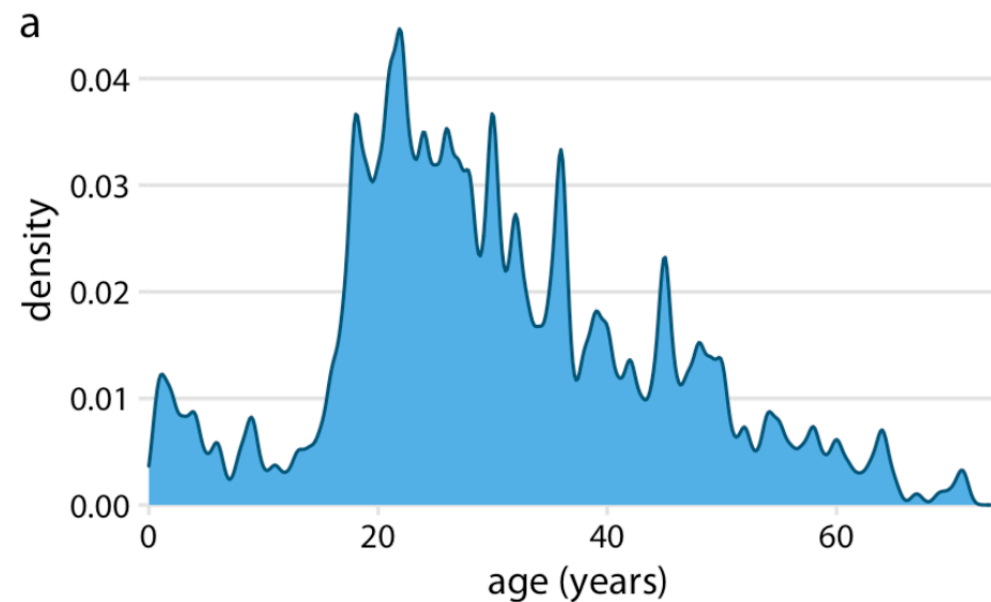
Different parabolic kernel widths (normalized)



KDEs are not immune to the binning problem

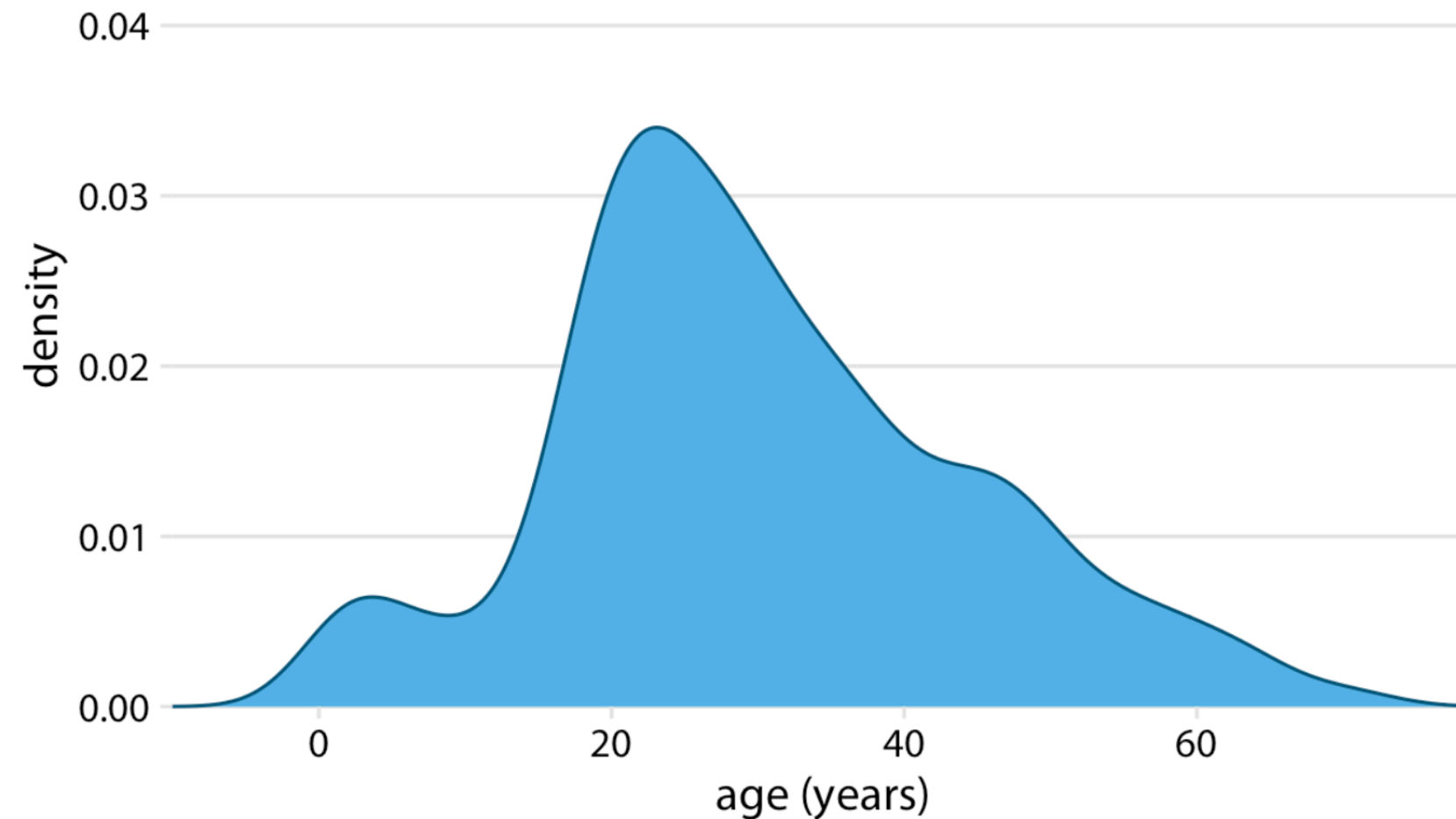


KDEs are not immune to the binning problem



- As number of data points gets large, shape of kernel (e.g. parabola) matters less.
- When number of data points are comparable to kernel size, KDEs are erratic and unreliable.

KDE boundary problem



- KDEs should only not extend beyond the data range (e.g. ages < 0). If it does, the situation is actually worse than it seems. Why?

KDEs and probability distributions

- KDEs describe a type of **empirical probability distribution**.
- Rather than fit the data to their own distribution, there are benefits to fitting it to a mathematically-specified (e.g. Gaussian) distribution.
 - Allows us to generalize better (i.e. if we had more data).
 - We can leverage its mathematical properties.
 - Example: Normalized Gaussian (e.g. symmetry, independence of mean and spread, area under curve).

Probability and random variables

- When we plot a probability distribution, the x-axis corresponds to the **random variable (RV)** and the y-axis shows the probability assigned to each value of the RV.
- Colored marble example.